Segmentez des clients d'un site e-commerce

Parcours Data Scientist| projet 5

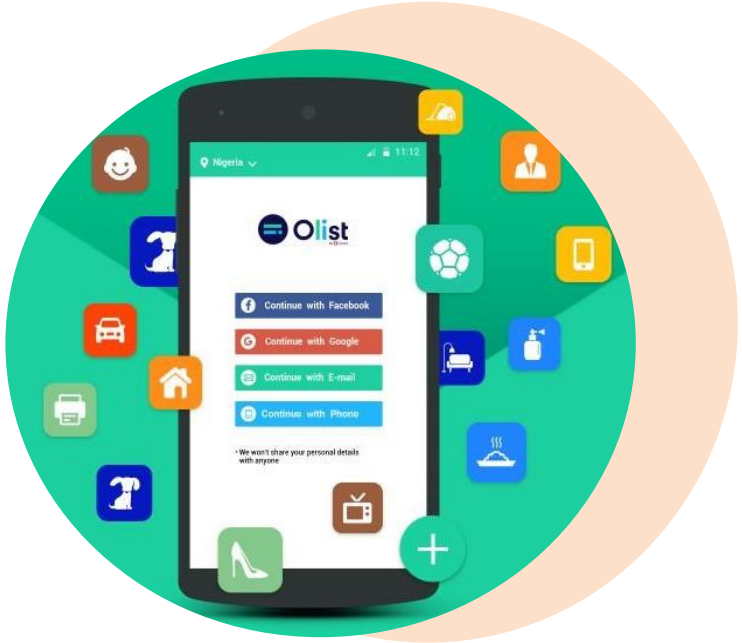
Rim BAHROUN

Février 2023

Segmentez des clients d'un site e-commerce

- **Problématique**

Olist: entreprise brésilienne de vente sur les marketplaces en ligne.



- **Mission**
  - fournir aux équipes d'e-commerce une description **actionnable** de la **segmentation des clients**.
  - fournir une proposition de **contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps.

Segmentez des clients d'un site e-commerce

1 Préparation du jeu de données

Extraire les données caractérisant les clients à partir de la base de données Olist.

Nettoyage, sélection et création de variables, analyse exploratoire …

2 Modélisation et segmentation des clients

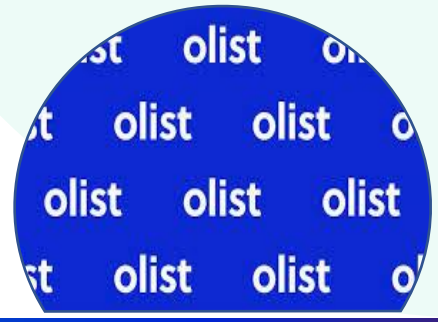
Segmenter les clients en fonction de leurs caractéristiques en utilisant les algorithmes de Machine Learning **non supervisés**.

Interpréter les segments obtenus d'un point de vue métier.

3 Simulation: contrat de maintenance

Analyser la stabilité temporelle de la segmentation pour évaluer une fréquence de maintenance

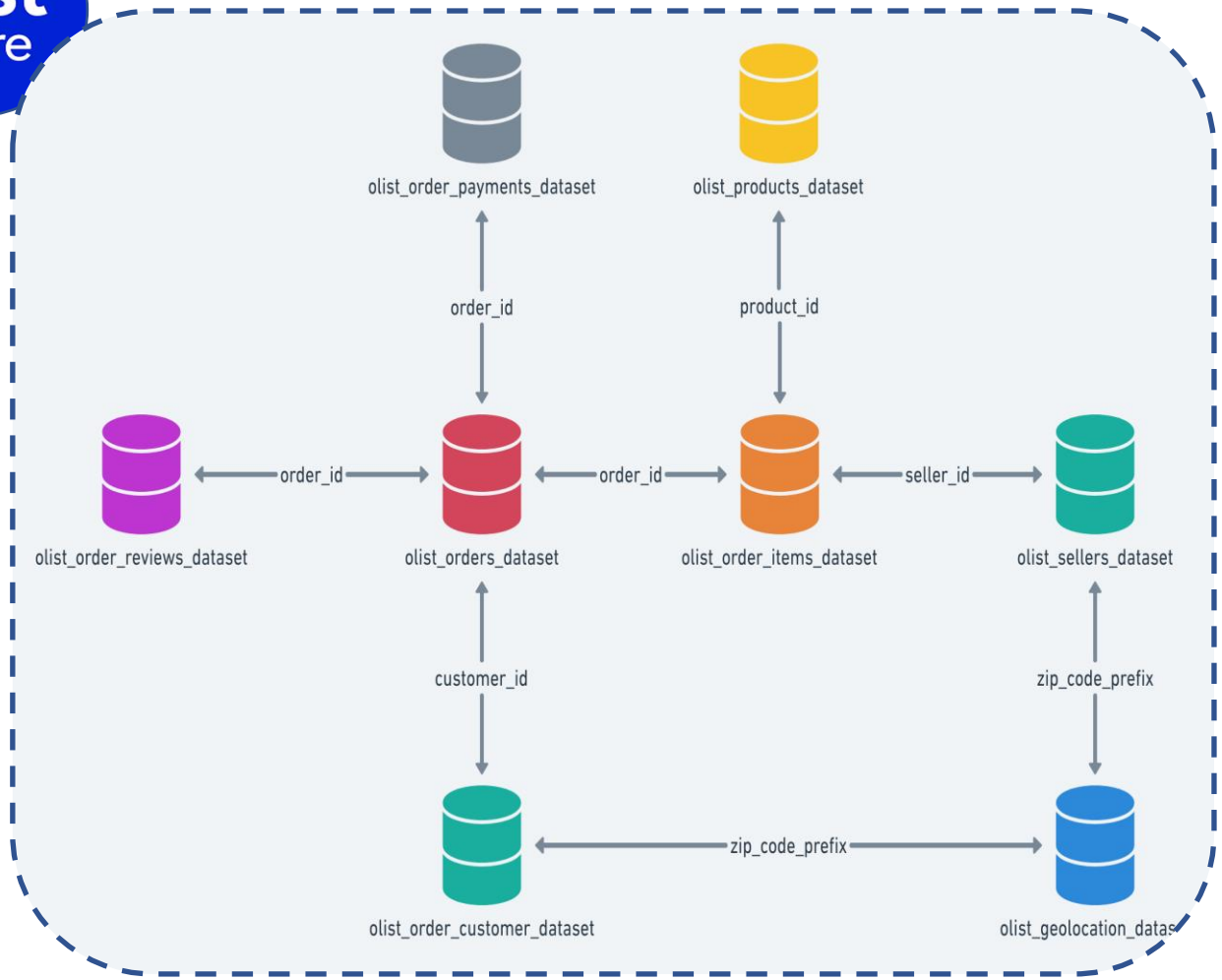
4 Conclusion



OPENCLASSROOMS

1. Préparation du jeu des données

Données à disposition : **9** fichiers **.csv**



Nettoyage et analyse exploratoire

- Jeu de données globalement **bien complété**
- Sélection des tables utiles pour la segmentation
- Jointure des tables (merge) selon les clés primaires
- Sélection des commandes **déjà livrées** uniquement
- Imputation des valeurs manquantes
- Correction des types de données
- **Sélection et création de nouvelles variables**
- Agrégation des données **par commande**
- Agrégation des données **par client**

1. Préparation du jeu des données

Feature engineering

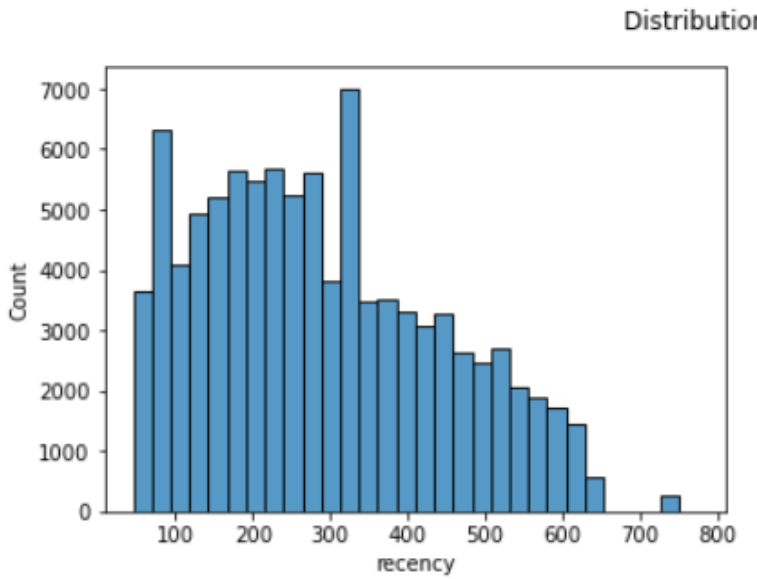
- Satisfaction
- Payement
- Livraison
- Articles
- Catégories
- Localisation
- Monetory
- Frequency
- Recency

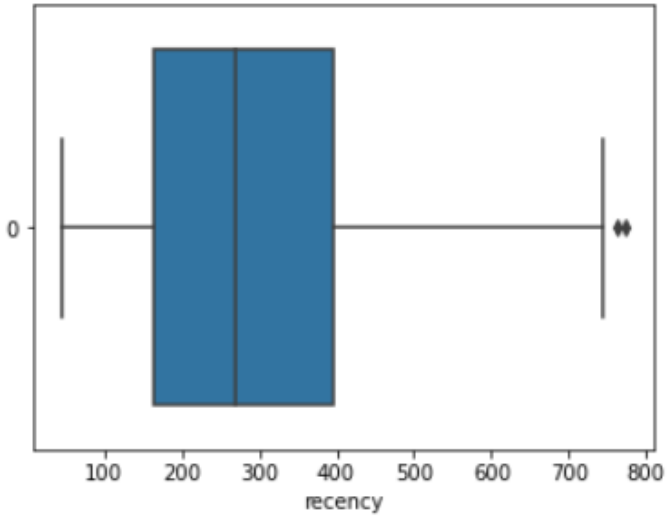
1. Préparation du jeu des données

Feature engineering et analyse exploratoire

Recency

La durée en jours entre une date de référence et la dernière date d'achat.





	recency
count	94983.000000
mean	288.334197
std	152.984601
min	45.000000
25%	164.000000
50%	269.000000
75%	397.000000
max	774.000000

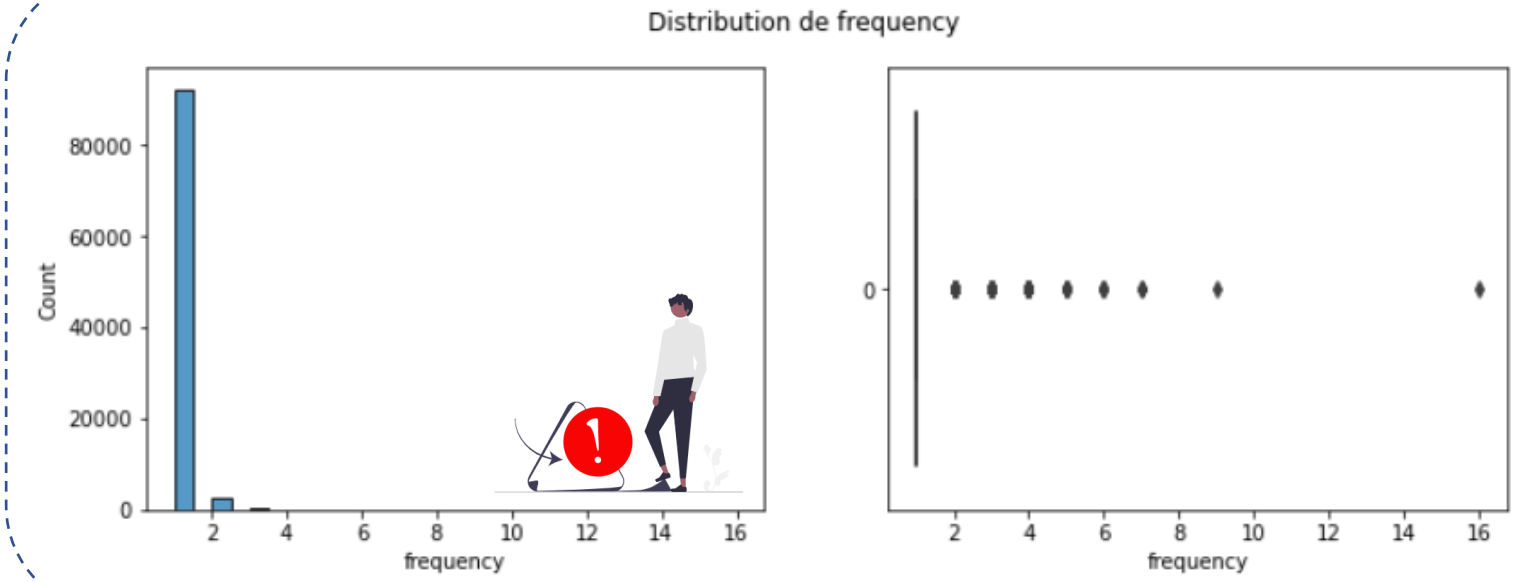
Les commandes remontent jusqu'à 2 ans

1. Préparation du jeu des données

Feature engineering et analyse exploratoire

Frequency

Le nombre fois qu'un client a commandé pendant la période d'étude.



	frequency
count	94983.000000
mean	1.033859
std	0.210811
min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
max	16.000000

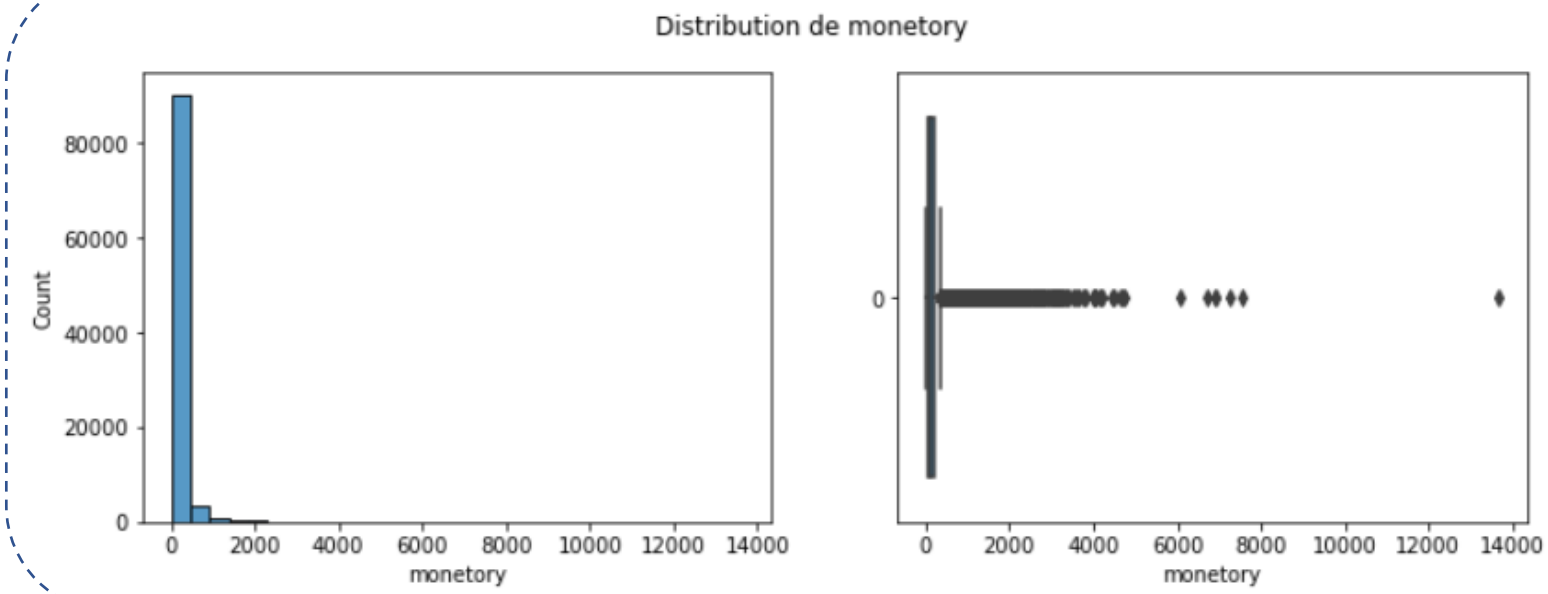
97% des clients ont commandé une seule fois

1. Préparation du jeu des données

Feature engineering et analyse exploratoire

Monetory

Le montant total des achats par client pendant la période d'étude.



	monetory
count	94983.000000
mean	165.696655
std	226.747246
min	9.590000
25%	63.100000
50%	107.900000
75%	182.945000
max	13664.080000

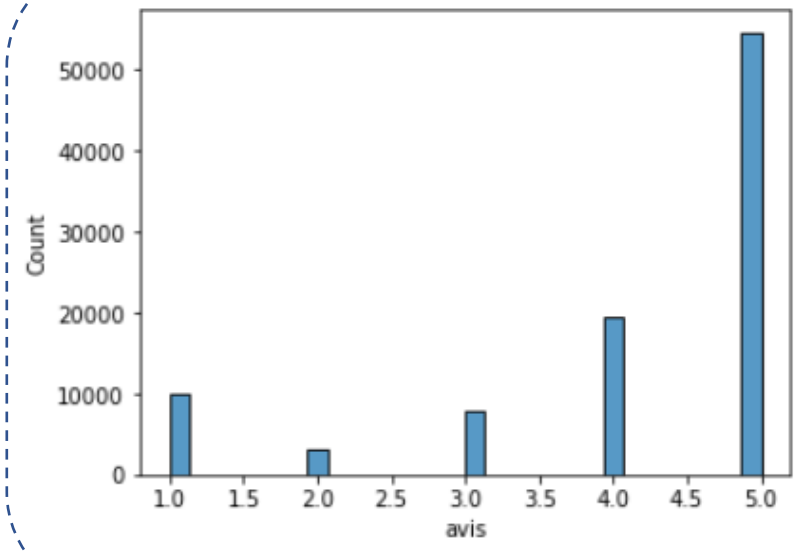
75% des clients ont payé moins de 200 Réal Brésilien

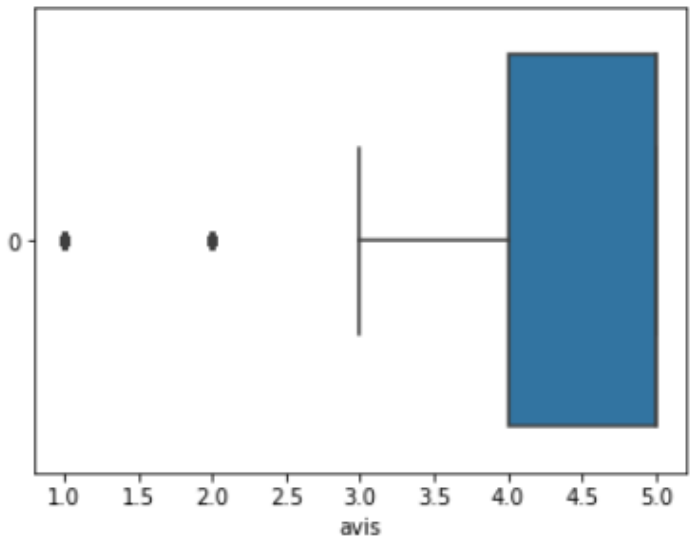
1. Préparation du jeu des données

Feature engineering et analyse exploratoire

Satisfaction

La moyenne des ‘review_score’ donnés par le client.





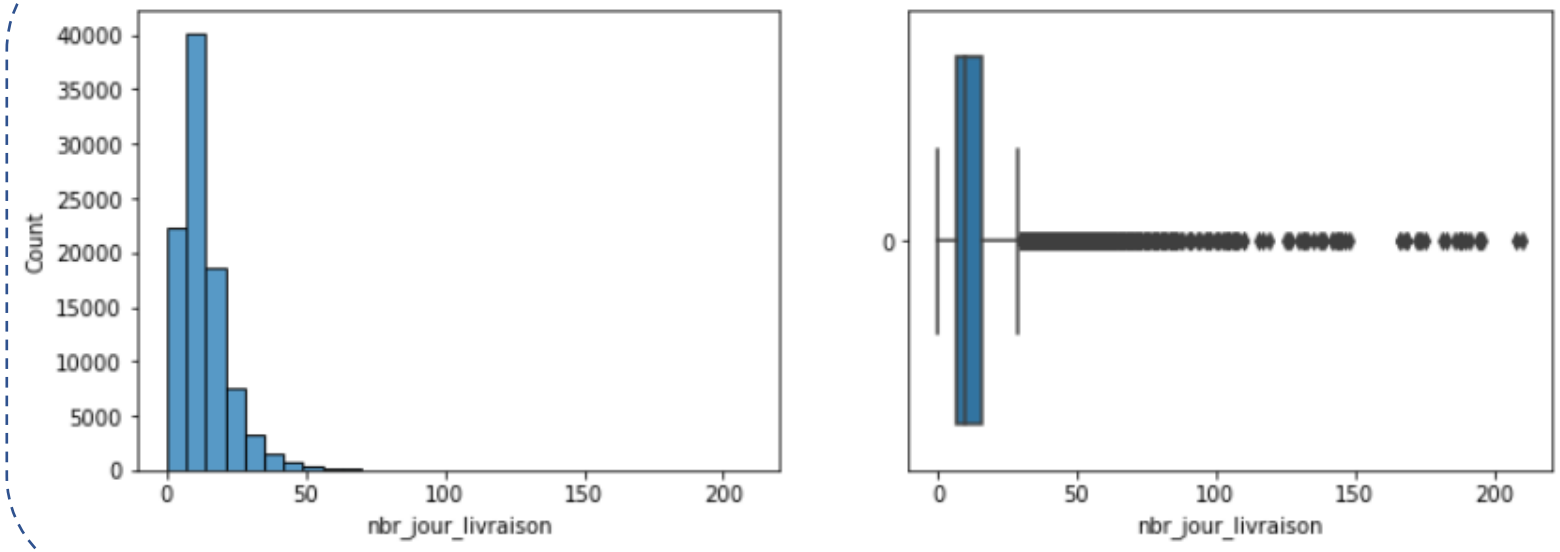
La majorité des clients sont satisfaits

1. Préparation du jeu des données

Feature engineering et analyse exploratoire

Livraison

Le nombre de jours entre la commande et la livraison.



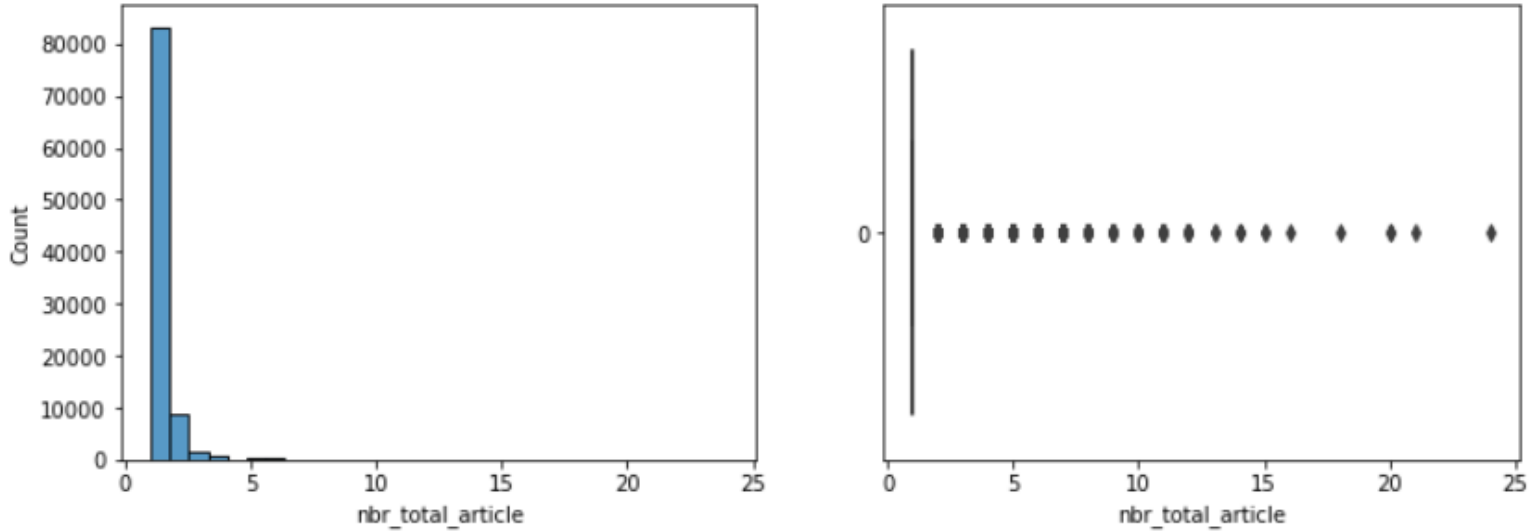
La majorité des livraisons se font dans les 30 jours

1. Préparation du jeu des données

Feature engineering et analyse exploratoire

Articles

Le nombre total d'articles achetés par un client.



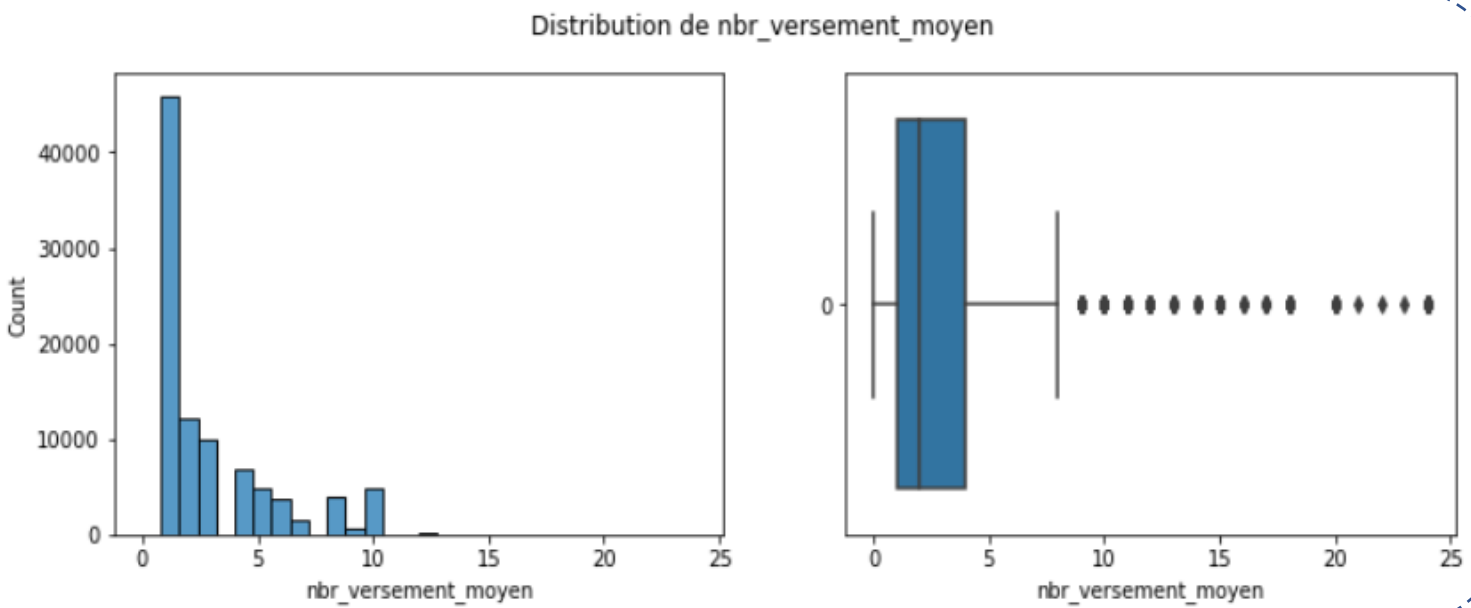
La majorité des clients ont acheté un seul article

1. Préparation du jeu des données

Feature engineering et analyse exploratoire

Payement

Le nombre d'échéances (1 si le payement est en une seule fois).



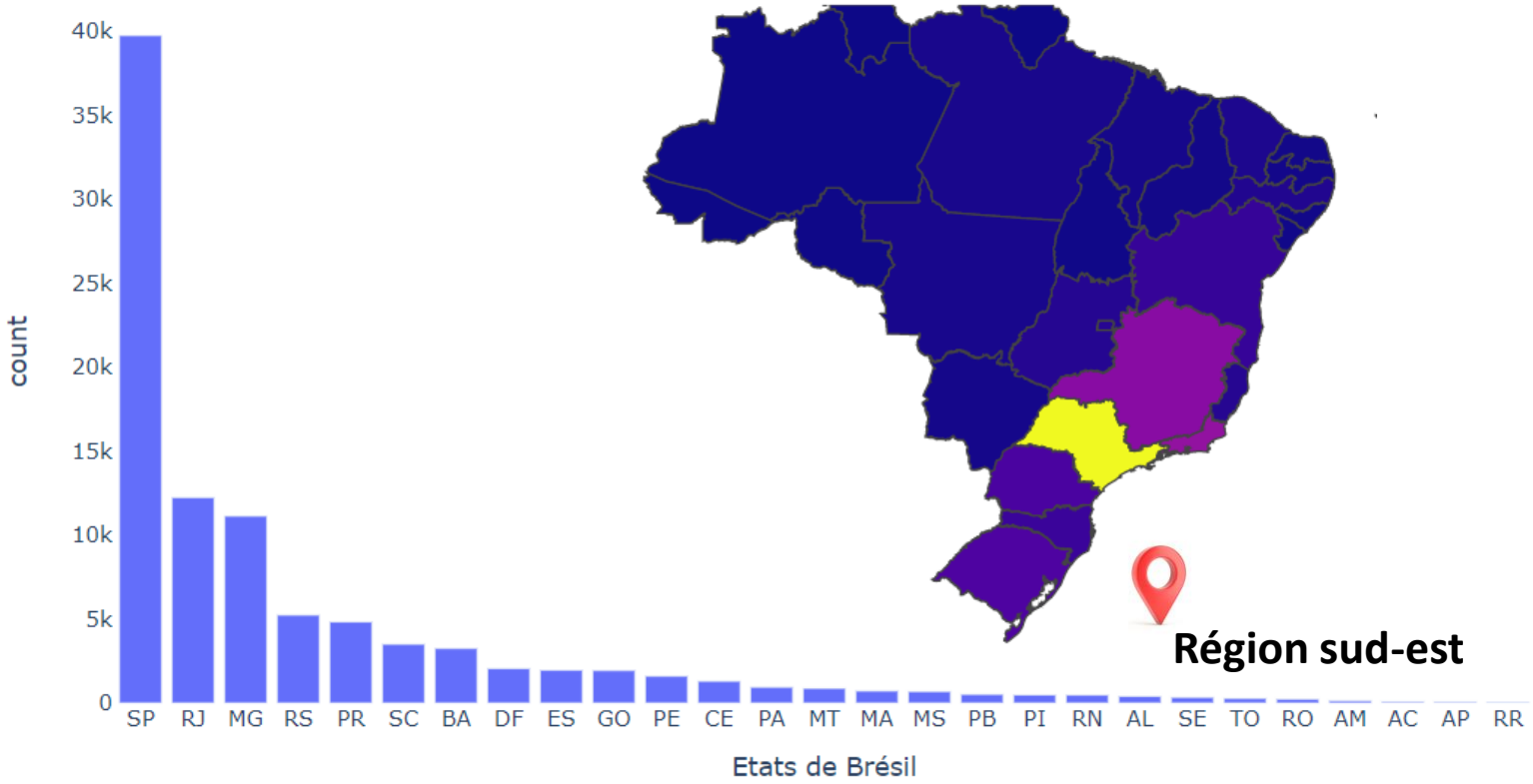
Une grande partie des clients ont payé en une seule fois

1. Préparation du jeu des données

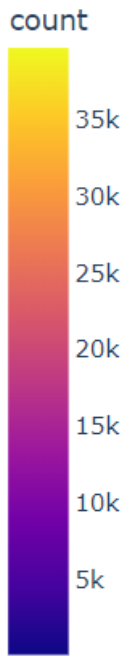
Feature engineering et analyse exploratoire

Localisation

L'Etat où habite le client.



Région sud-est



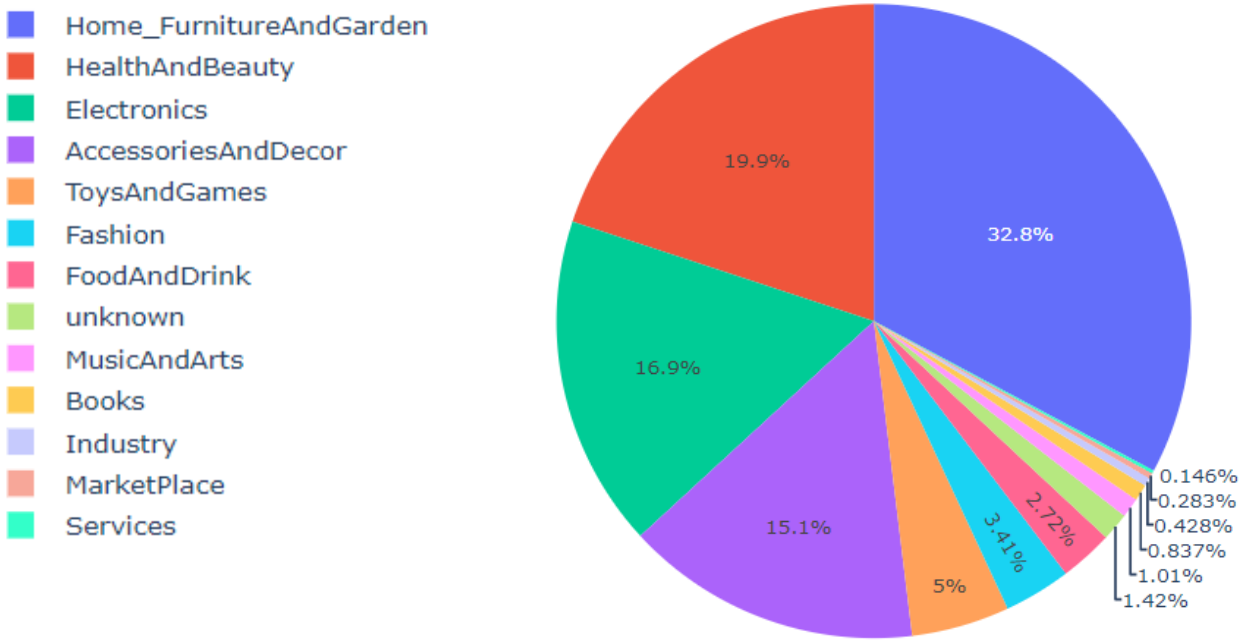
Les clients sont localisés principalement en région sud-est

1. Préparation du jeu des données

Feature engineering et analyse exploratoire

Catégories

La catégorie la plus achetée par le client parmi un regroupement de 12 catégories.



La catégorie la plus achetée est les articles de maison

1. Préparation du jeu des données

Synthèse

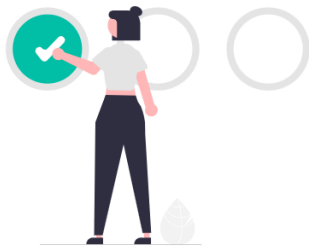
Dans cette première partie, nous avons effectué :

- le nettoyage et l'imputation des données
- la jointure et l'agrégation des données par commande et par client
- la sélection et la création **de 7 variables pertinentes** pour la segmentation des clients: **recency, frequency, monetory, satisfaction, livraison, articles et payement**
- **l'analyse exploratoire** de ces variables
- les données de type **catégorielles ne seront pas retenues** pour la segmentation

Ces opérations ont permis de préparer le jeu de donnée pour l'étape de modélisation

Jeu final agrégé par client
7 variables
94983 clients

#	Column	Non-Null Count	Dtype
0	recency	94983 non-null	int64
1	frequency	94983 non-null	int64
2	monetory	94983 non-null	float64
3	avis	94983 non-null	int64
4	nbr_jour_livraison	94983 non-null	int64
5	nbr_versement_moyen	94983 non-null	int64
6	nbr_total_article	94983 non-null	int64



Segmentez des clients d'un site e-commerce

1 Préparation du jeu de données

Extraire les données caractérisant les clients à partir de la base de données Olist.

Nettoyage, sélection et création de variables, analyse exploratoire …

2 Modélisation et segmentation des clients

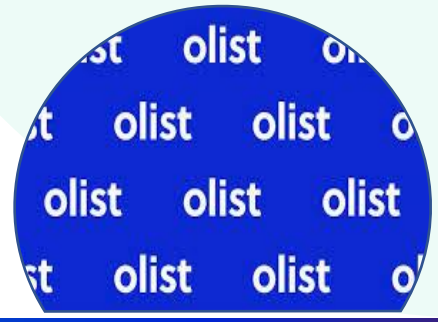
Segmenter les clients en fonction de leurs caractéristiques en utilisant les algorithmes de Machine Learning **non supervisés**.

Interpréter les segments obtenus d'un point de vue métier.

3 Simulation: contrat de maintenance

Analyser la stabilité temporelle de la segmentation pour évaluer une fréquence de maintenance

4 Conclusion



OPENCLASSROOMS

2. Modélisation et présentation des résultats

Modélisation

Jeu de données nettoyé

Prétraitement des données

Standardisation des variables numériques **StandardScaler**

X: données clients 7 variables

Modélisation

- Implémentation des **modèles de classification non supervisé**
- Choix du nombre de segments/ Adaptation des hyperparamètres
- Interprétation des clusters du point de vue métier

Evaluations des performances

- Comparaison des modèles
- Visualisation des clusters
- Choix du modèle final pour la segmentation

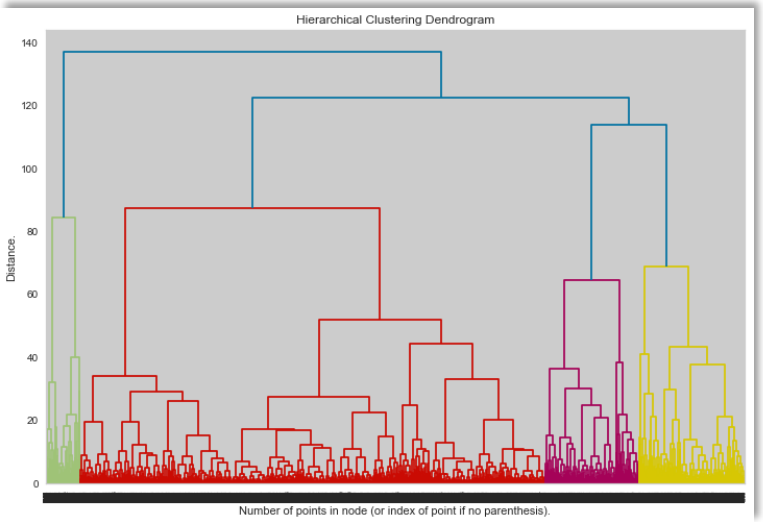
Description **actionnable** de la **segmentation des clients**.

2. Modélisation et présentation des résultats

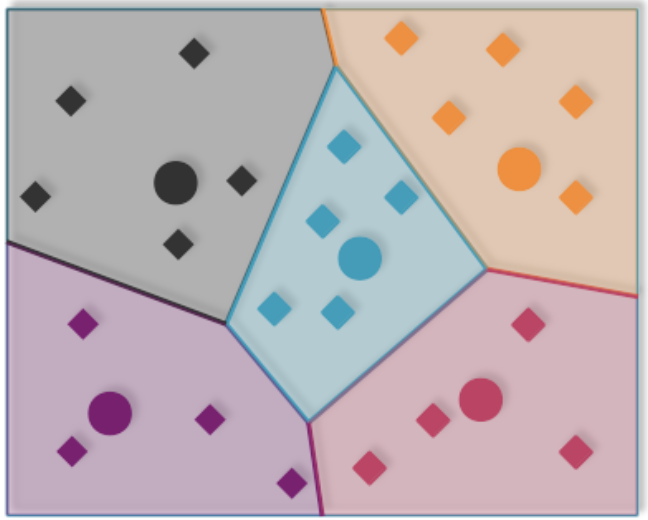
Modélisation

Algorithmes de clustering non supervisé

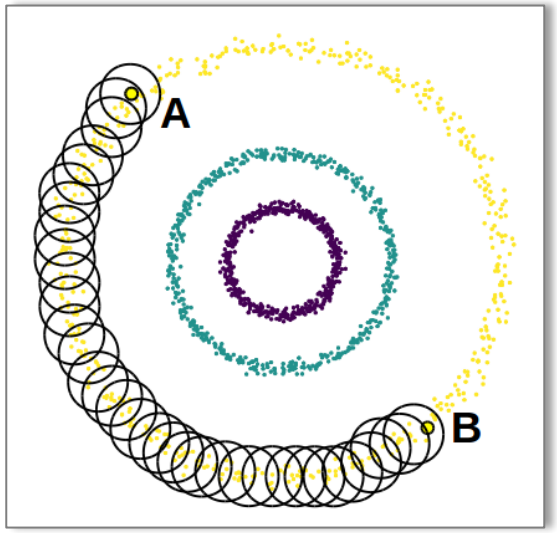
Clustering hiérarchique



K-means



DBSCAN



2. Modélisation et présentation des résultats

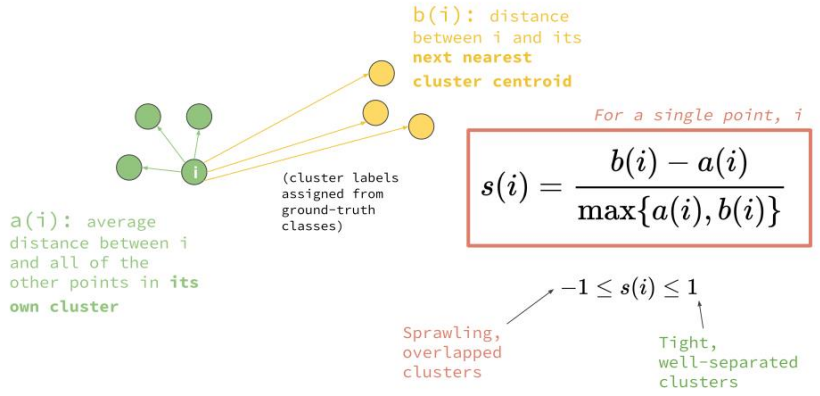
Modélisation

Silhouette score

Pour chaque point, son coefficient de silhouette est **la différence entre** la distance moyenne avec **les points du même groupe** que lui (a(i) **cohésion**) et la distance moyenne **avec les points des autres groupes voisins** (b(i) **séparation**). Il est entre -1 et 1.

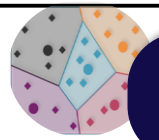
Si cette différence est **négative**, le point est en moyenne plus proche du groupe voisin que du sien : il est donc **mal classé**. À l'inverse, si cette différence est **positive**, le point est en moyenne plus proche de son groupe que du groupe voisin : il est donc **bien classé**.

Le coefficient de silhouette proprement dit est la moyenne du coefficient de silhouette pour tous les points.



2. Modélisation et présentation des résultats

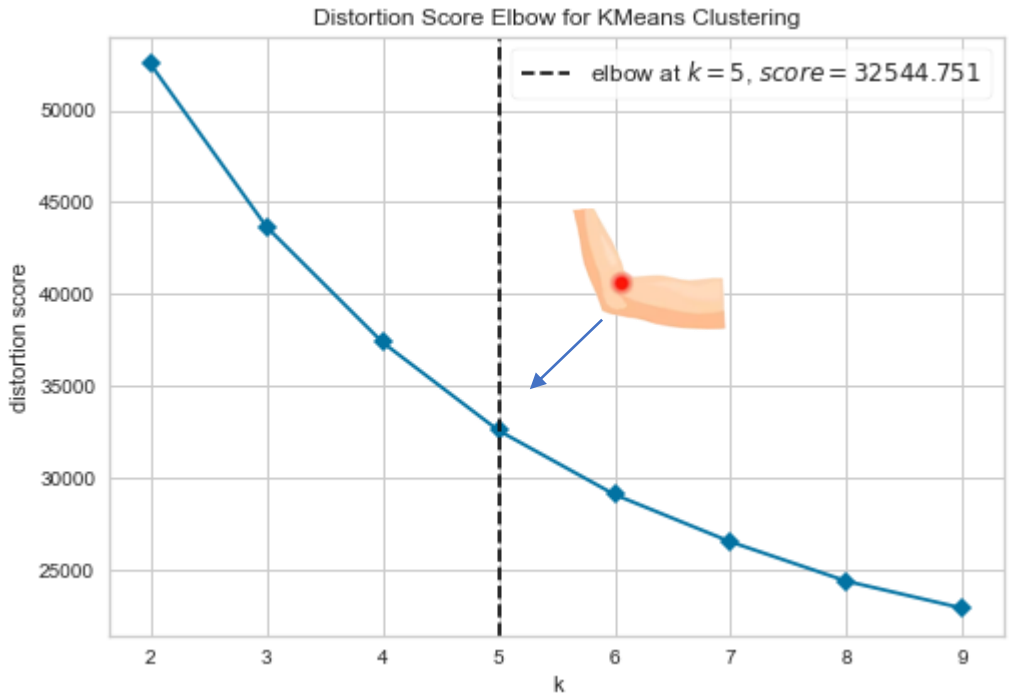
Modélisation



K-means

Choix du nombre de clusters

K = 5

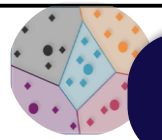


Distortion score: la somme des carrées des distances de chaque point à son centre de classe assignée.

Modélisation sur 10% des données

2. Modélisation et présentation des résultats

Modélisation



K-means

Stabilité de l'algorithme à l'initialisation

Modèle stable

K-means initialisé avec **les composantes PCA:** méthode déterministe

➡ annulation des effets aléatoires de l'initialisation des centroïdes.

```
1  # Create a k-means clustering model. Initialisation des centroïdes avec pca
2  pca_km = PCA(n_components=5).fit(X_scaled)
3  kmeans = KMeans(init=pca_km.components_, n_clusters=5, n_init=1)
```

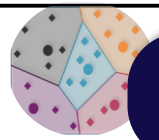
Scores de stabilité à l'initialisation

Iteration	FitTime	ARI
Iter 0	0.044s	1.000
Iter 1	0.033s	1.000
Iter 2	0.034s	1.000
Iter 3	0.034s	1.000
Iter 4	0.034s	1.000
Iter 5	0.035s	1.000
Iter 6	0.036s	1.000
Iter 7	0.035s	1.000
Iter 8	0.035s	1.000
Iter 9	0.035s	1.000

Modélisation sur 10% des données

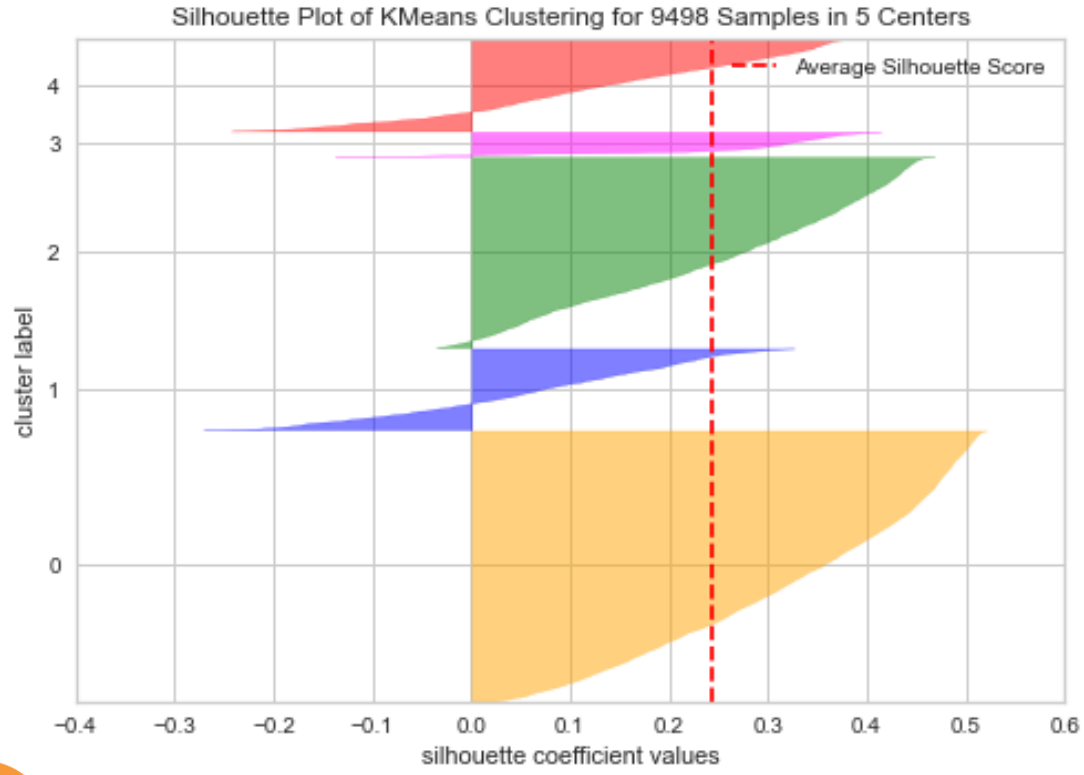
2. Modélisation et présentation des résultats

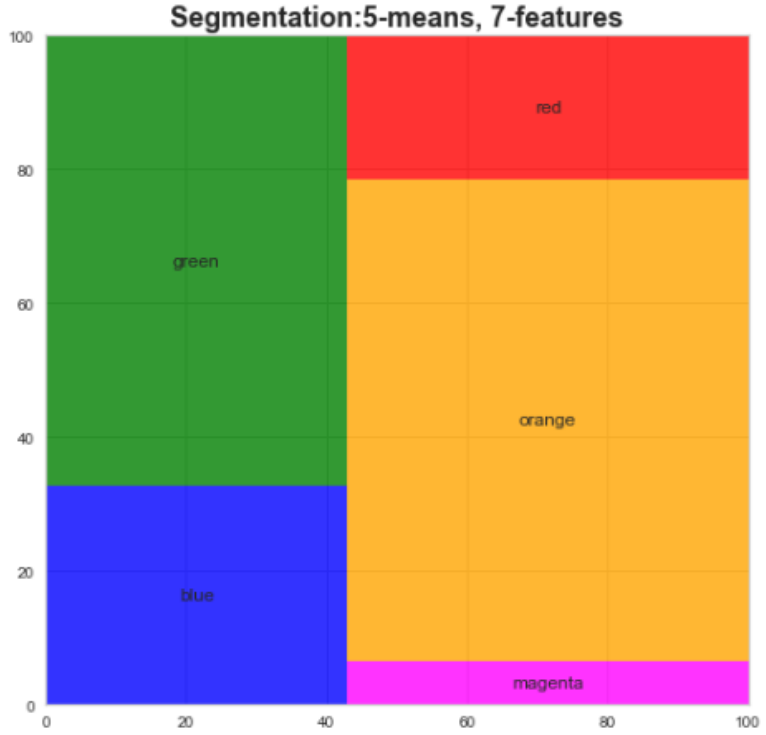
Modélisation



K-means

Silhouette score et répartition des clusters





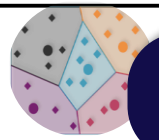
Silhouette score = 0,24

Temps d'exécution = 0,47 s

Modélisation sur 10% des données

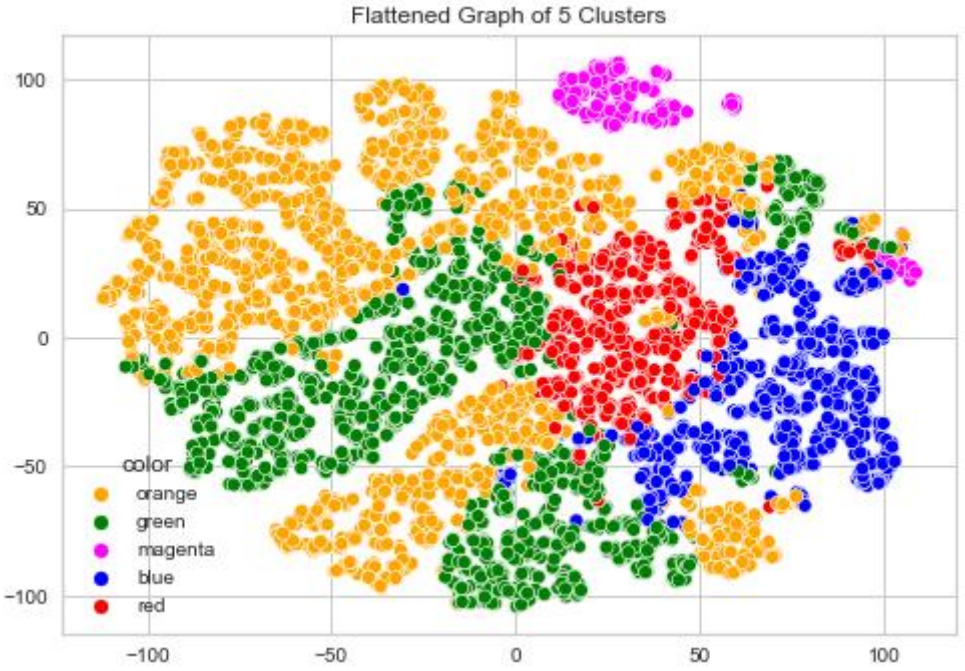
2. Modélisation et présentation des résultats

Modélisation

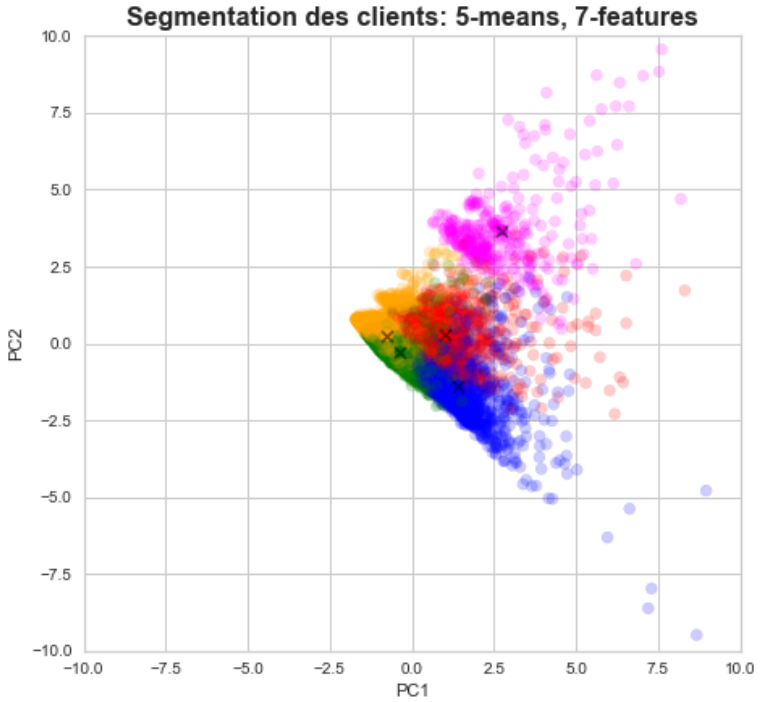


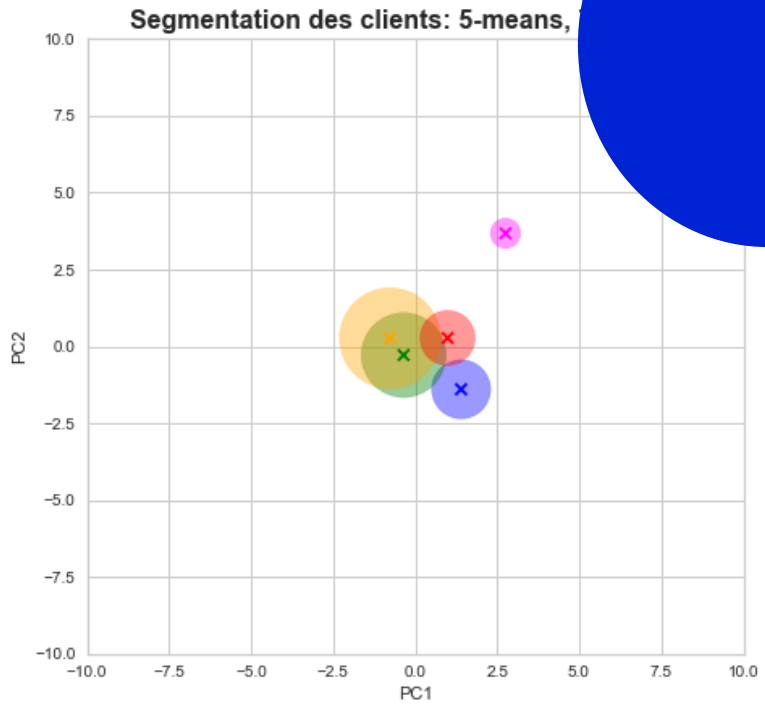
K-means

Projection des données segmentées dans un plan 2D



TSNE





PCA – 2D

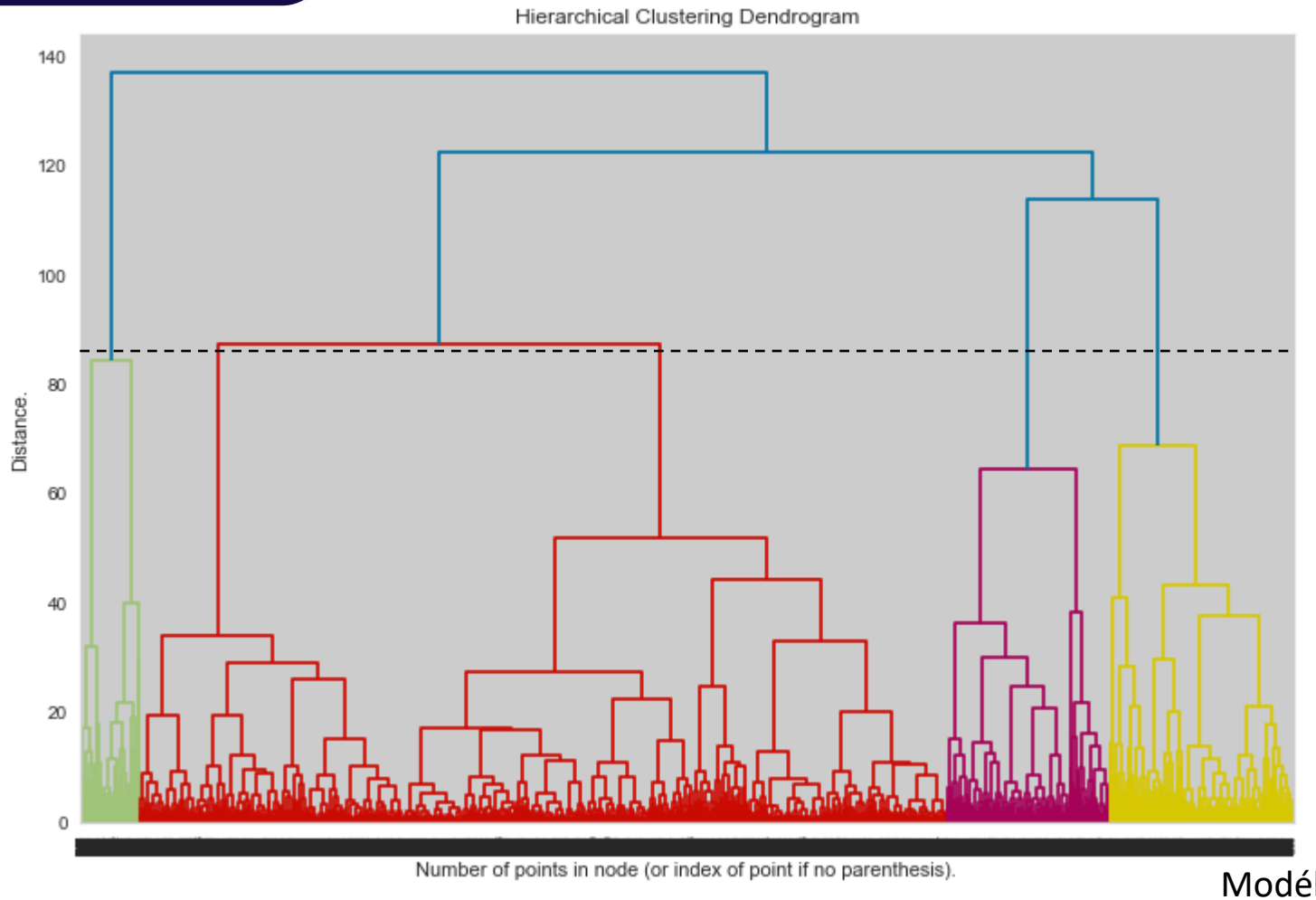
(0,43 variance expliquée)

Modélisation sur 10% des données

2. Modélisation et présentation des résultats

Modélisation

Clustering agglomératif



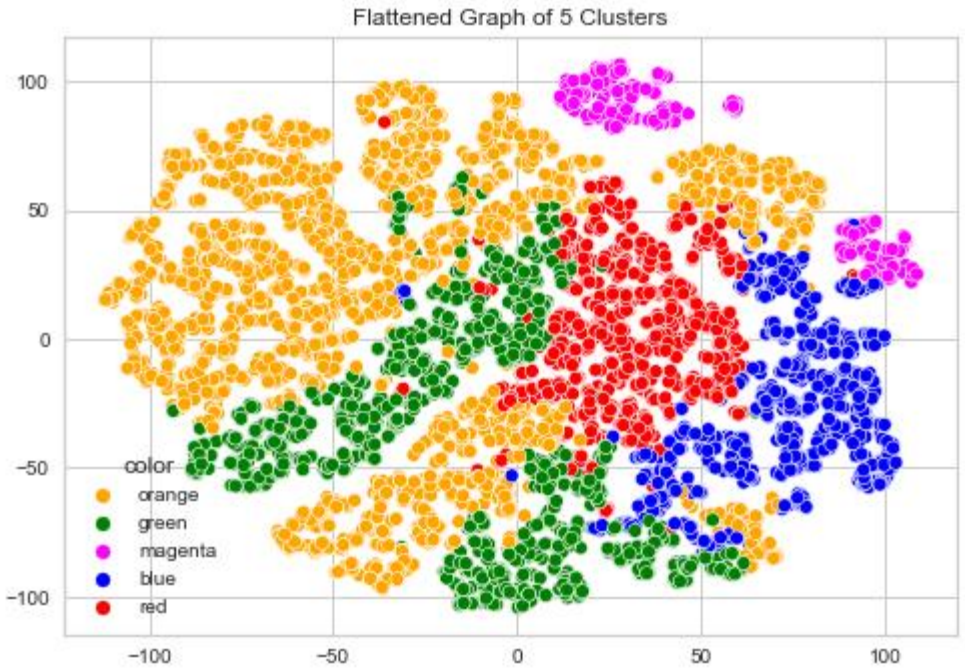
Modélisation sur 10% des données

2. Modélisation et présentation des résultats

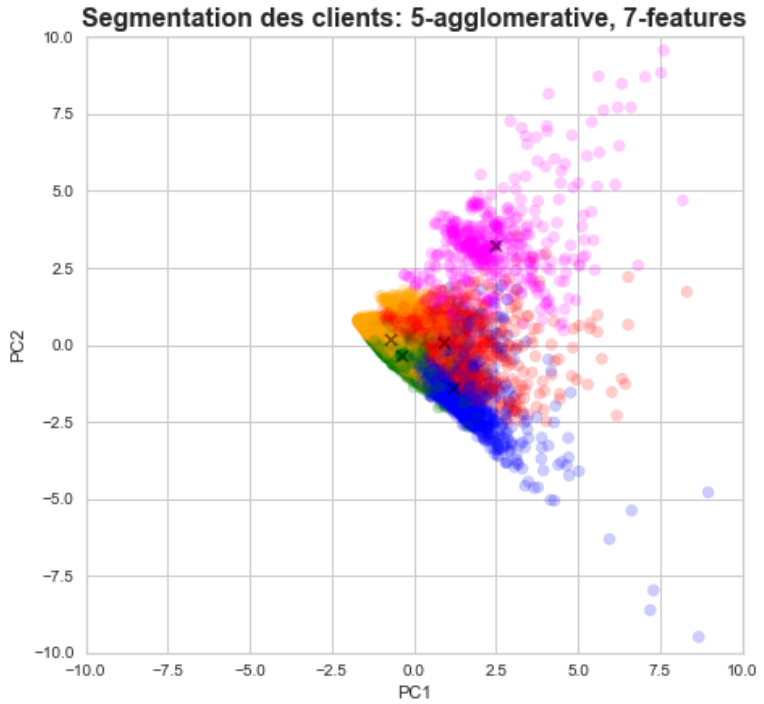
Modélisation

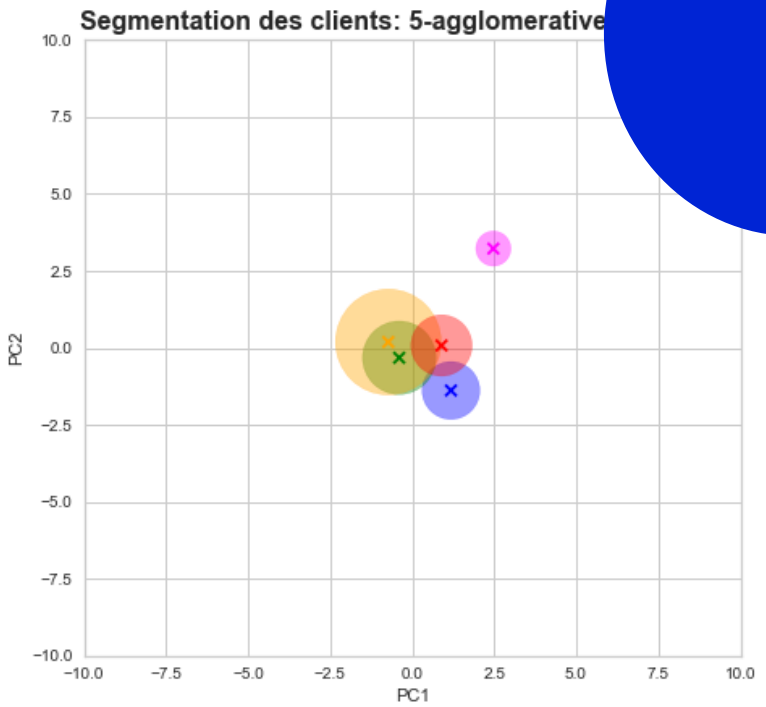
Clustering agglomératif

Projection des données segmentées dans un plan 2D



TSNE





PCA – 2D (0,43 variance expliquée)

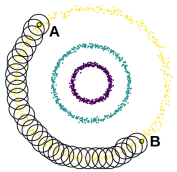
Silhouette score = 0,2

Temps d'exécution = 5,15 s

Modélisation sur 10% des données

2. Modélisation et présentation des résultats

Modélisation



DBSCAN

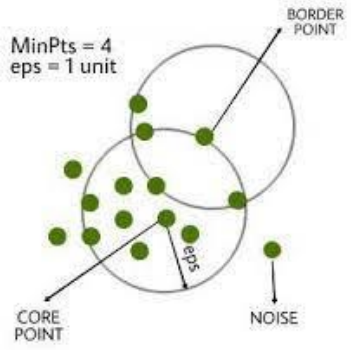
Choix des hyperparamètres

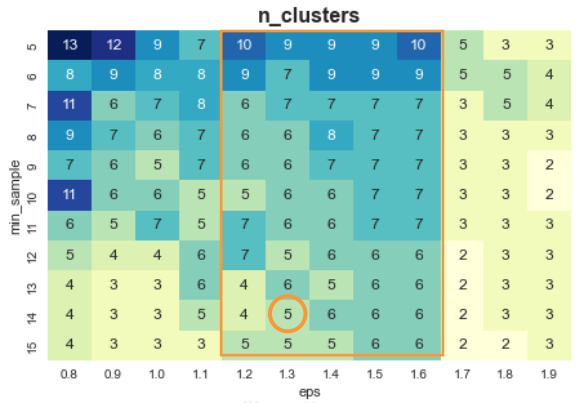
DBSCAN requiert deux hyperparamètres:

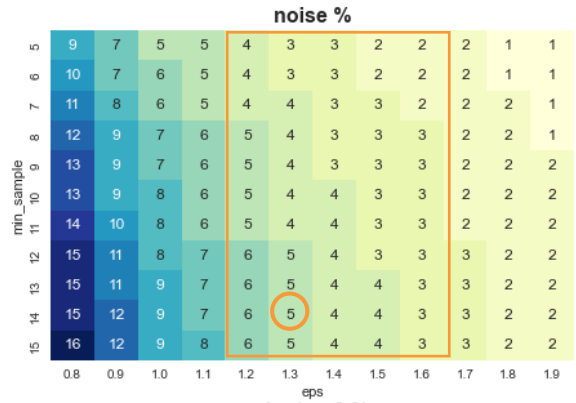
eps (ε) la taille du voisinage (le rayon de la boule),

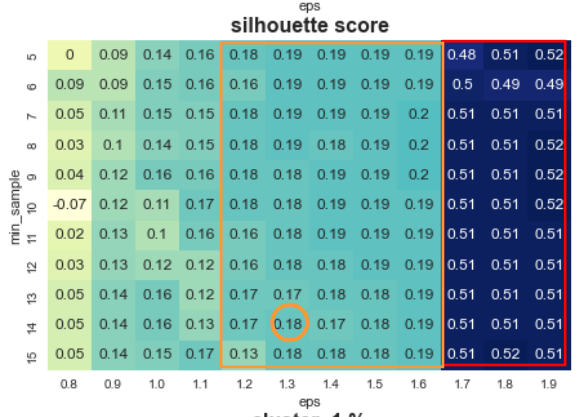
min_sample la densité minimale à dépasser (le nombre de voisins dans la boule pour être considéré un point interieur).

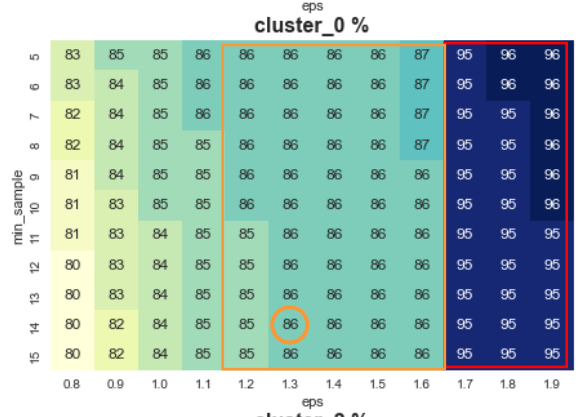
Choix difficile :
Min_sample = 14 (soit 2x7 features)
Eps = 1,3 (5 clusters)

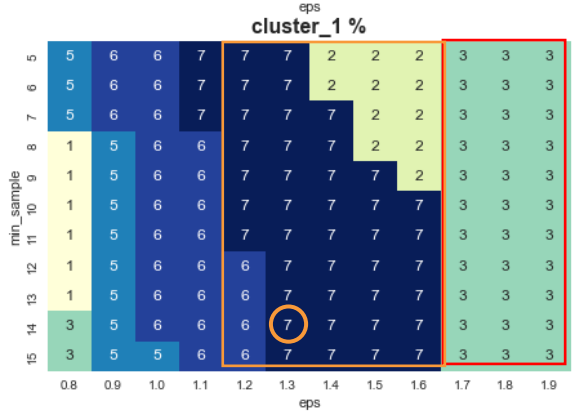


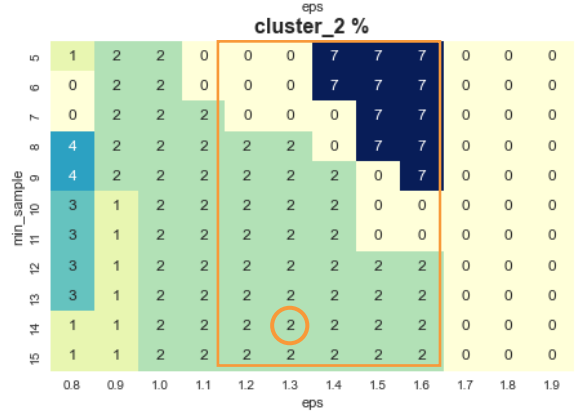






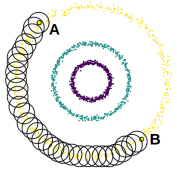






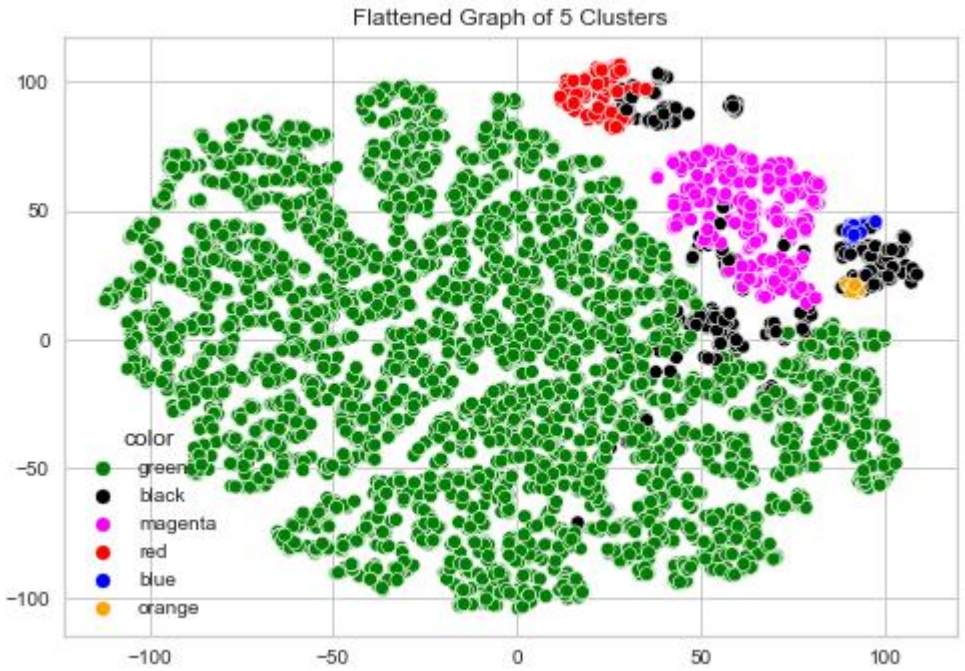
2. Modélisation et présentation des résultats

Modélisation

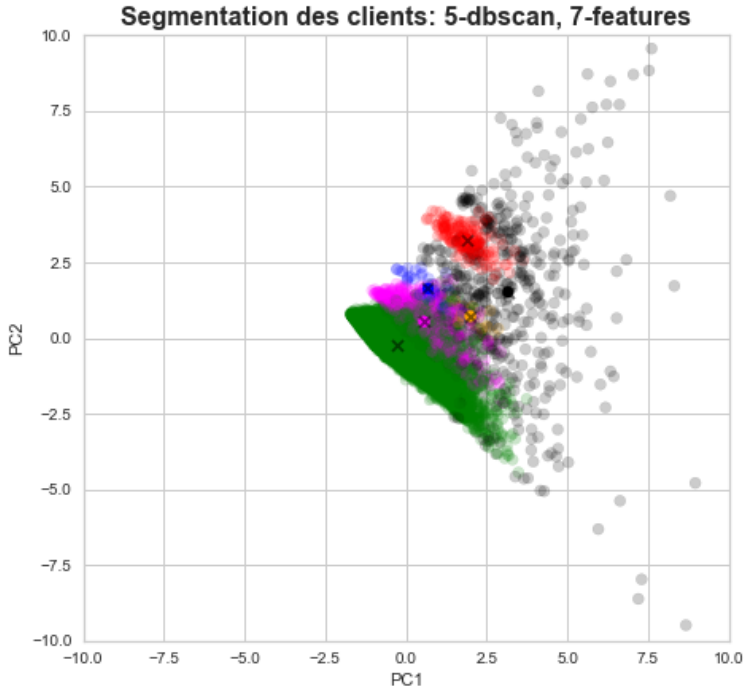


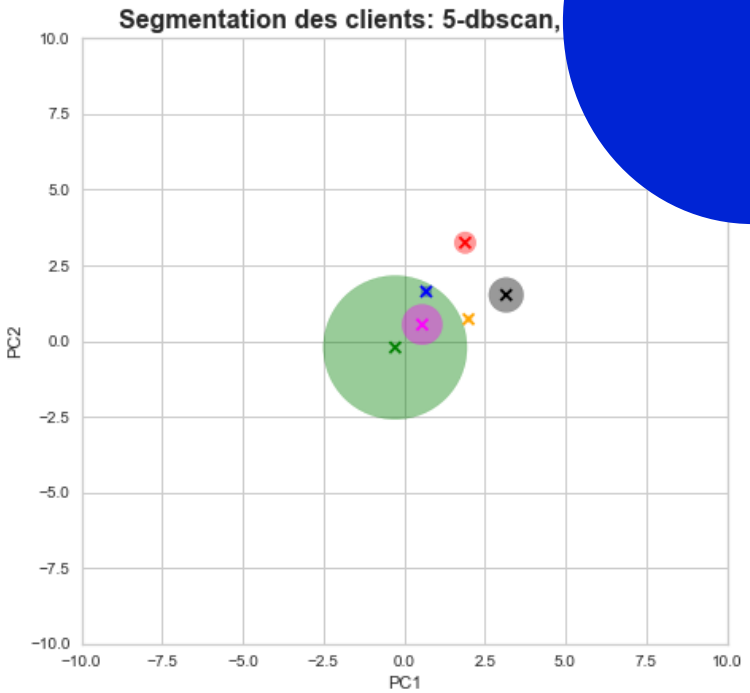
DBSCAN

Projection des données segmentées dans un plan 2D



TSNE





PCA – 2D (0,43 variance expliquée)

Silhouette score = 0,18

Temps d'exécution = 2,14 s

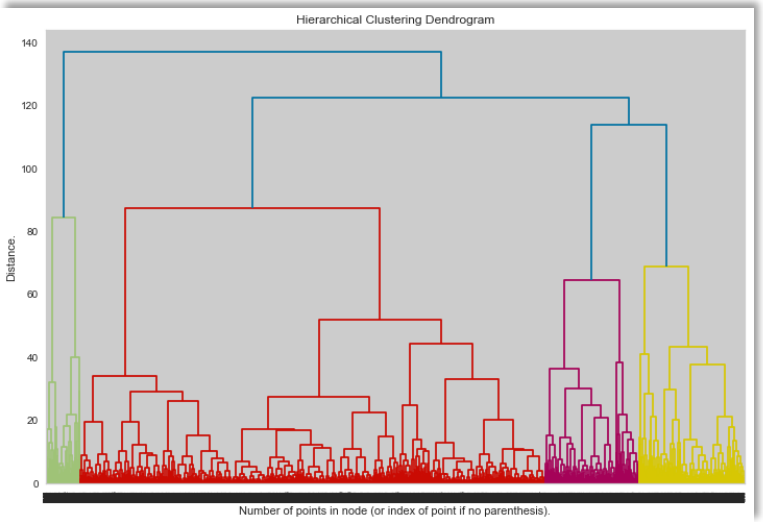
Modélisation sur 10% des données

2. Modélisation et présentation des résultats

Comparaison des algorithmes

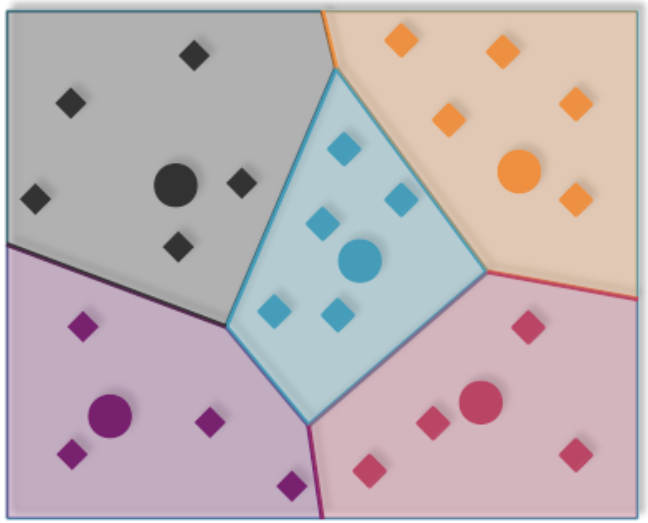
Algorithmes de clustering non supervisé

Clustering hiérarchique



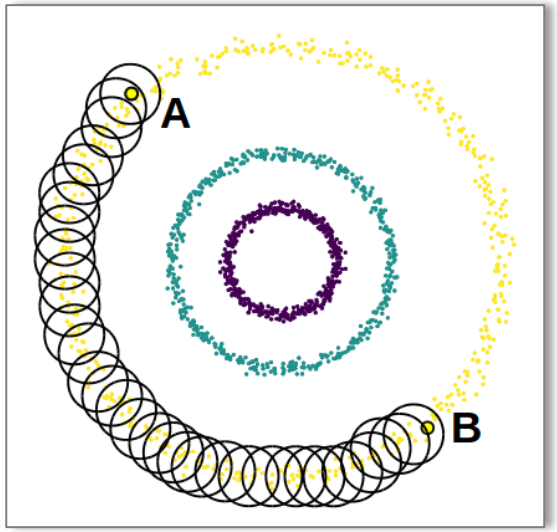
Flexibilité : nombre de clusters, distances utilisées
Complexité algorithmique lourde !
Temps de calcul et espace de mémoire importants

K-means



Efficace en temps de calcul
Le nombre de clusters est donné à l'avance !
Clusters convexes !

DBSCAN

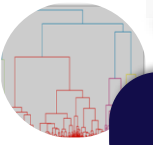


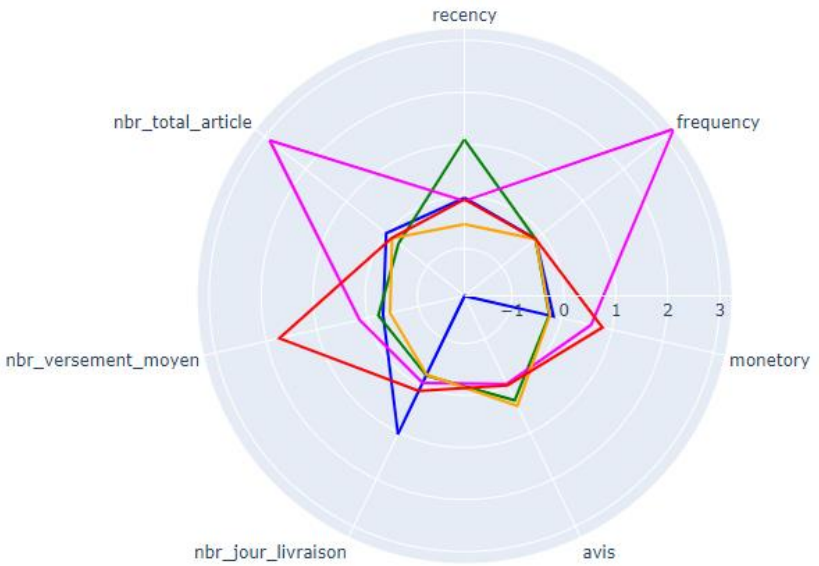
Flexibilité : nombre de clusters, forme arbitraire
Clusters de densité comparable
Choix délicat des hyperparamètres: eps, n_min

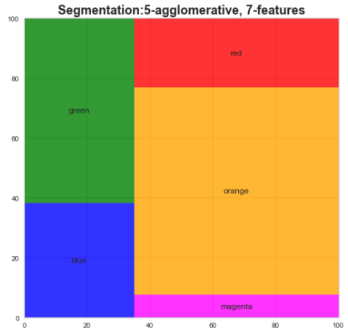
2. Modélisation et présentation des résultats

Comparaison des algorithmes

Clustering agglomératif

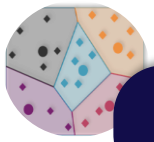


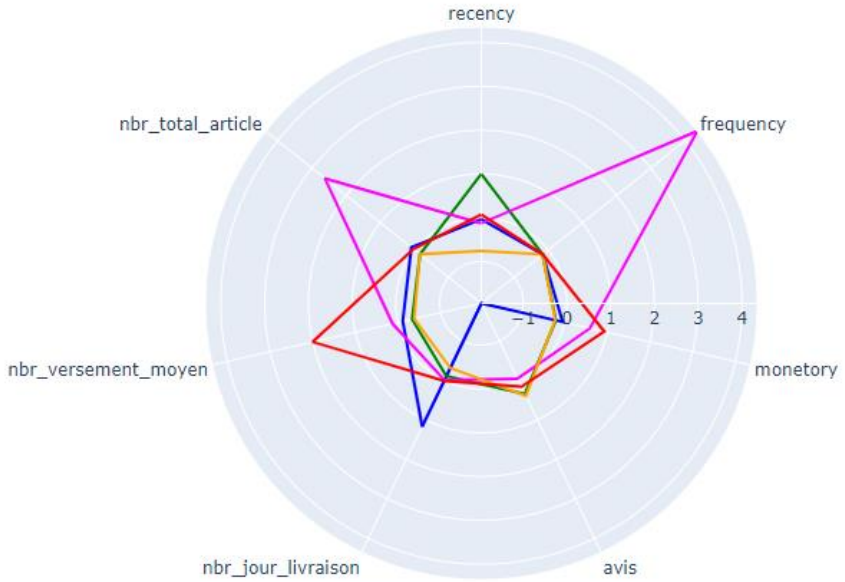


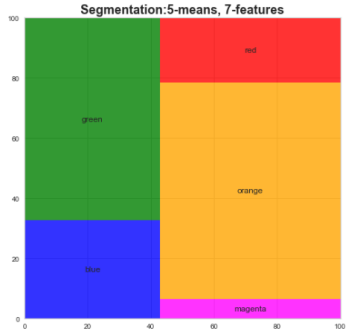


ARI score : 0,58 entre les deux segmentations

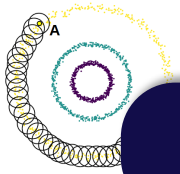
K-means

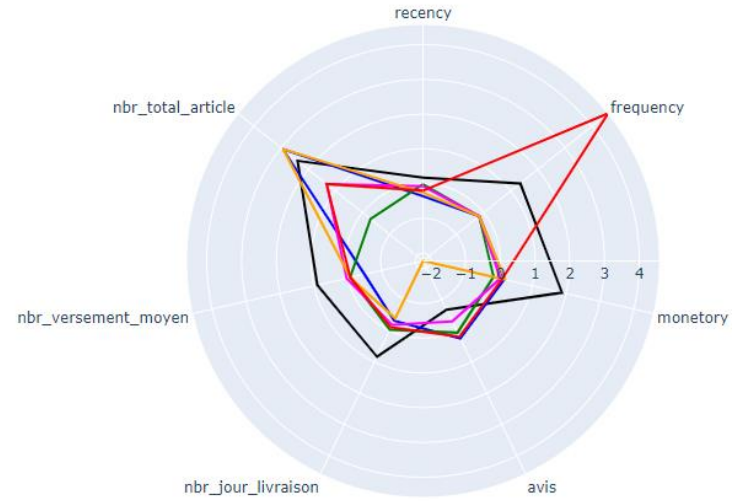


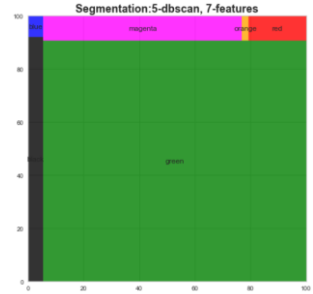


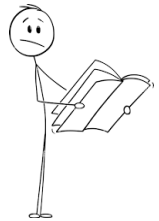


DBSCAN



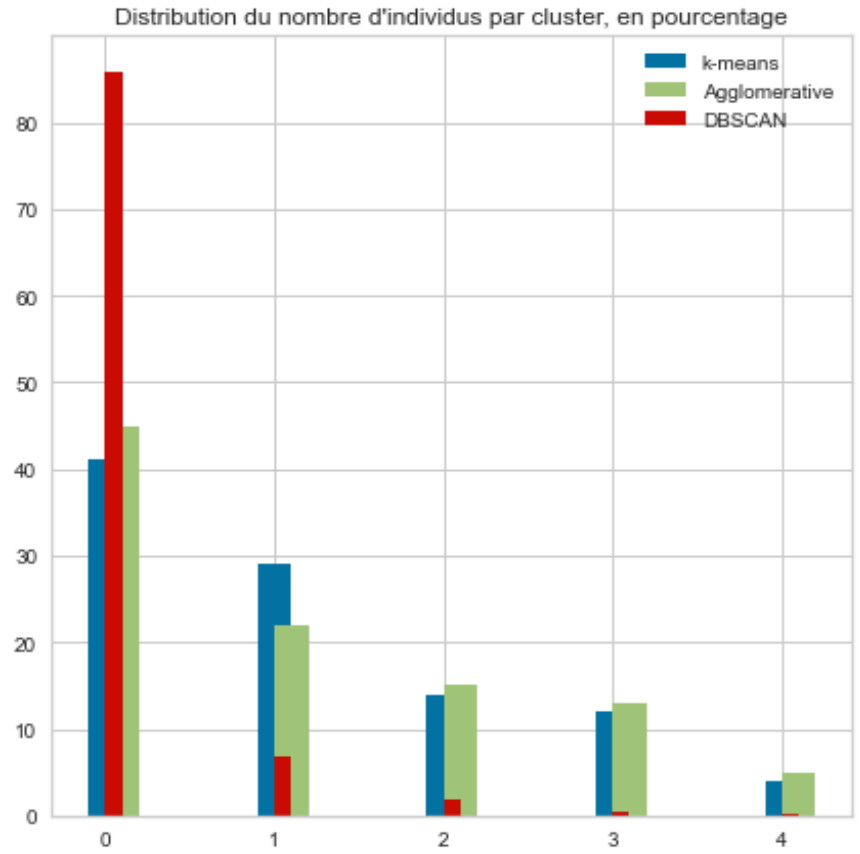




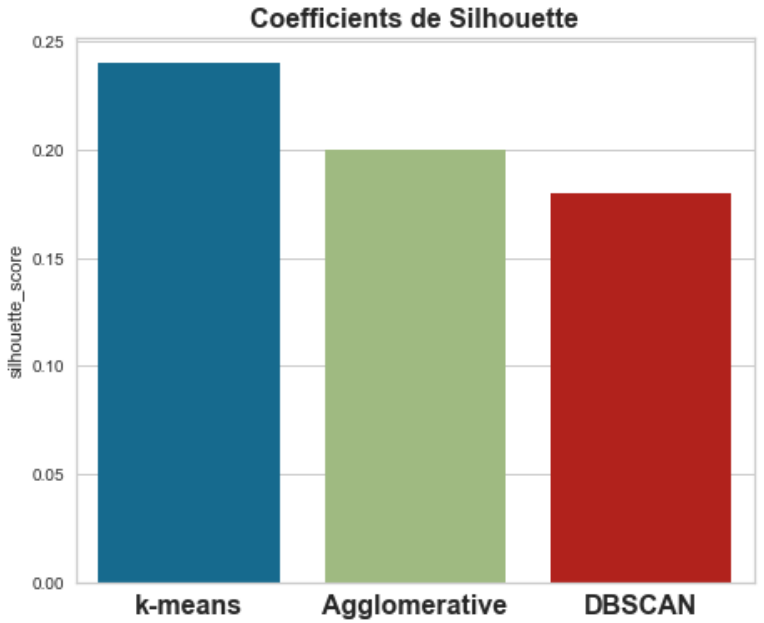


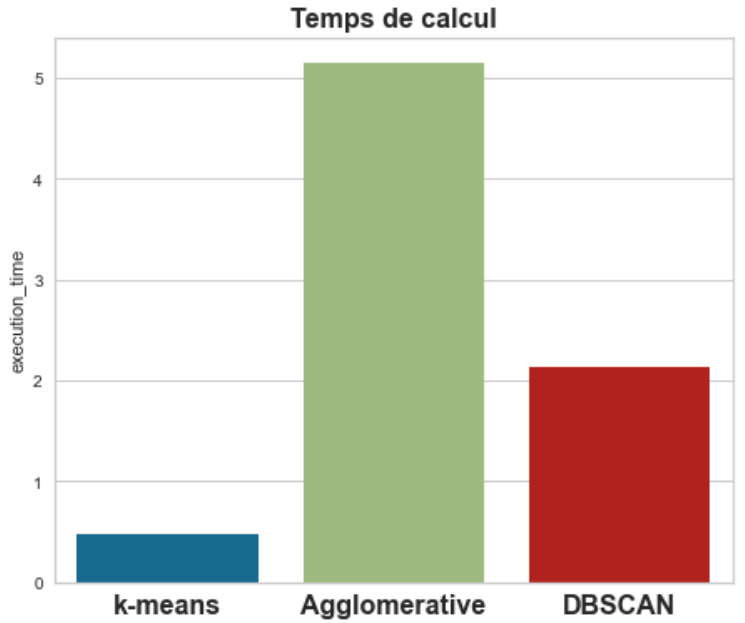
Modélisation sur 10% des données

2. Modélisation et présentation des résultats



Comparaison des algorithmes

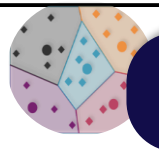




Modèle final retenu: k-means 5 clusters

2. Modélisation et présentation des résultats

Modélisation

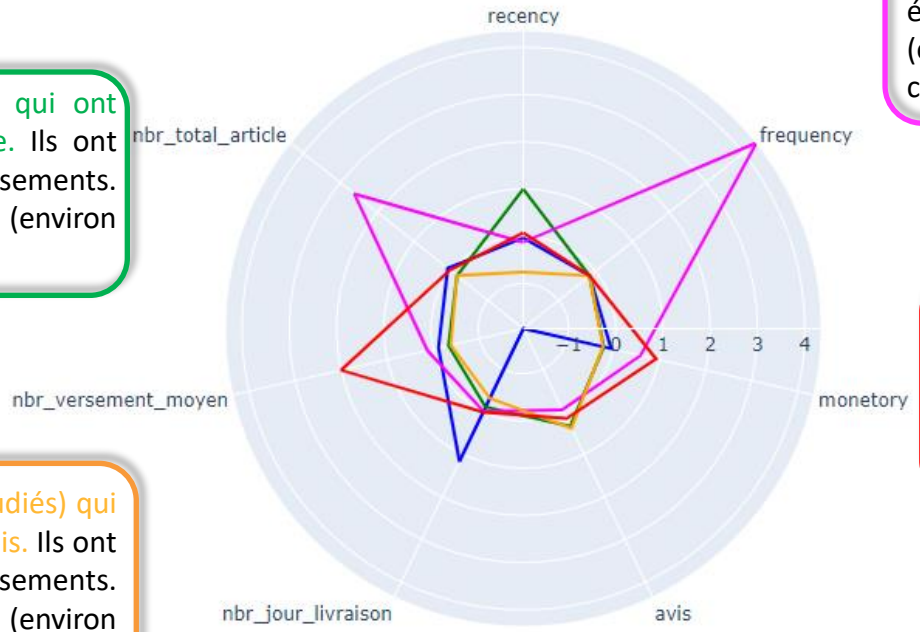


K-means

Description actionnable des 5 clusters

Churned (les infidèles): (29% des clients étudiés) qui ont acheté en moyenne 1 fois il y a plus d'une année. Ils ont acheté un seul article, et effectué en moyenne 2 versements. Ils ont payé une petite somme de 119 Réal brésilien (environ 20 euros). Ils étaient très satisfaits.

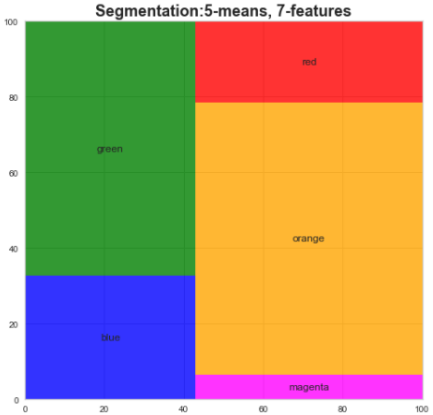
New customers (les nouveaux): (41% des clients étudiés) qui ont acheté en moyenne 1 fois dans les derniers 6 mois. Ils ont acheté un seul article, et effectué en moyenne 2 versements. Ils ont payé une petite somme de 115 Réal brésilien (environ 20 euros). Ils étaient très satisfaits.



color	recency	frequency	monetory	avis	nbr_jour_livraison	nbr_versement_moyen	nbr_total_article
blue	284	1	154	2	24	3	1
green	441	1	119	5	12	2	1
magenta	269	2	299	4	13	3	3
orange	172	1	115	5	10	2	1
red	300	1	376	4	13	8	1

Loyal (les fidèles): (4% des clients étudiés) qui ont acheté plus qu'une fois et plus d'un article cette année. Ils ont dépensé également une bonne somme en moyenne 299 Réal brésilien (environ 55 euros). Ils ont payé en moyenne sur 3 fois. C'est clients sont fidèles et sont généralement satisfaits.

Big spenders (les dépensiers): (12% des clients étudiés) qui ont dépensé une somme importante en moyenne 374 Réal brésilien (environ 67 euros). Ils ont payé en moyenne sur 8 fois. Ils ont acheté en moyenne une seule fois et un seul article la dernière année. Ils étaient généralement satisfaits.



Unsatisfied (les détracteurs): (14% des clients étudiés) qui n'étaient pas du tout satisfaits en partie à cause de leur grand délais de livraison. Ils ont acheté en moyenne une fois un seul article dans cette année. Ils ont dépensé en moyenne 153 Réal brésilien (environ 27 euros).

Modélisation sur 10% des données

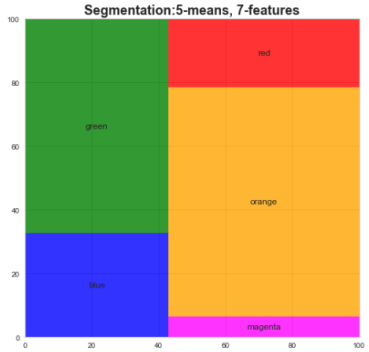
2. Modélisation et présentation des résultats

Synthèse

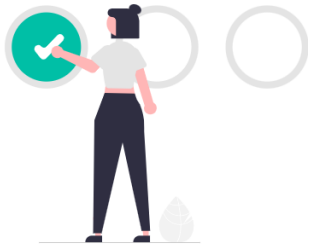
Dans cette deuxième partie, nous avons effectué :

- la segmentation non supervisée des clients par les algorithmes **k-means, agglomératif** et **DBSCAN**
- le choix du nombre de clusters adapté à notre problème métier: **5 clusters**
- le choix des hyperparamètres à partir de l'étude de plusieurs combinaison
- la comparaison entre les résultats de segmentation des algorithmes
- le choix du modèle final: **l'algorithme k-means avec 5 clusters**

Ces opérations ont permis de fournir une description actionnable des 5 segments de clients



1. Les dépensiers 12%
2. Les fidèles 4%
3. Les détracteurs 14%
4. Les nouveaux 41%
5. Les infidèles 29%



Segmentez des clients d'un site e-commerce

1 Préparation du jeu de données

Extraire les données caractérisant les clients à partir de la base de données Olist.

Nettoyage, sélection et création de variables, analyse exploratoire …

2 Modélisation et segmentation des clients

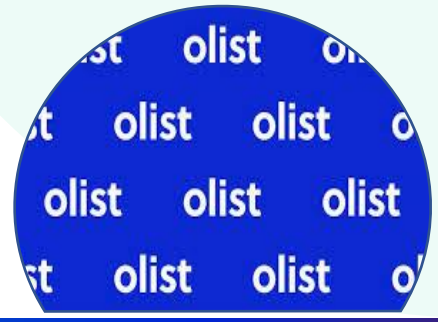
Segmenter les clients en fonction de leurs caractéristiques en utilisant les algorithmes de Machine Learning **non supervisés**.

Interpréter les segments obtenus d'un point de vue métier.

3 Simulation: contrat de maintenance

Analyser la stabilité temporelle de la segmentation pour évaluer une fréquence de maintenance

4 Conclusion



OPENCLASSROOMS

3. Contrat de maintenance

Méthodologie

Objectif: établir un contrat de maintenance de l'algorithme de segmentation client.

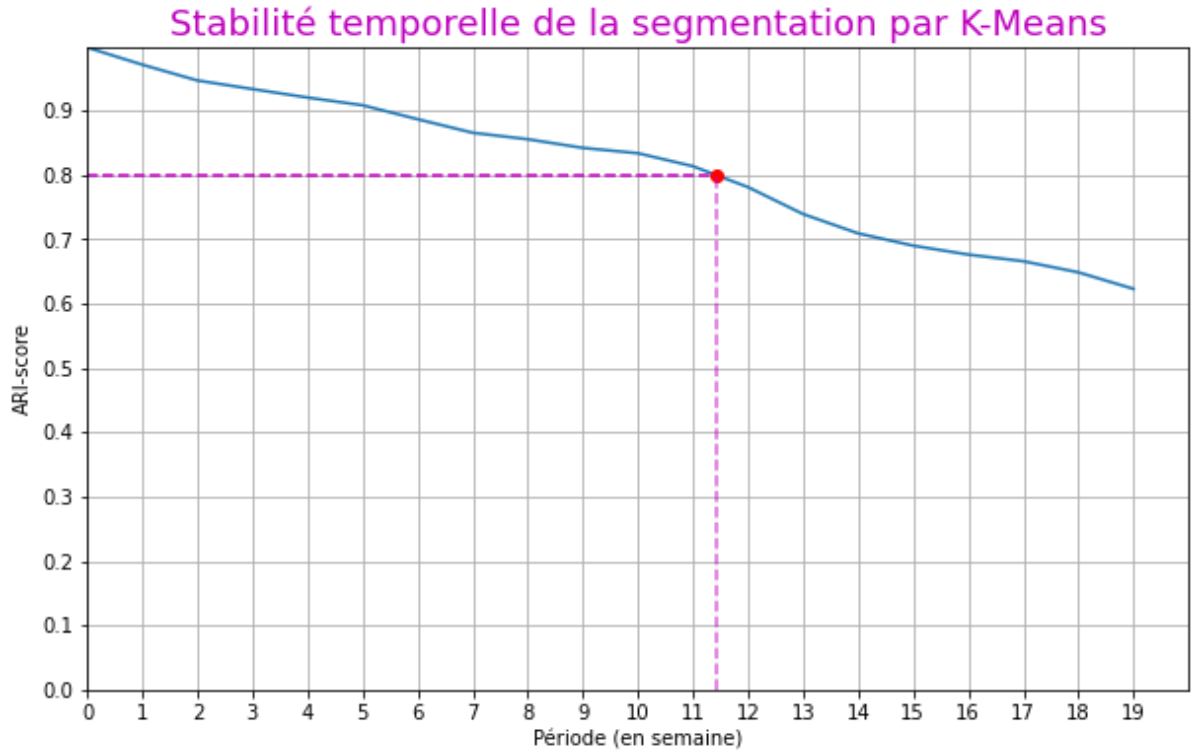
Le jeu de données initial Olist s'étend sur 2 ans. Une **période initiale** de **1an et demi** est donc fixée et un **modèle M0** est entrainé sur cette période.

Chaque semaine i, les nouvelles commandes sont prises en compte et toutes les variables utiles sont recalculées sur cette nouvelle période.

Un score ARI est calculé entre la segmentation des clients avec :
- le **modèle M0** (entrainé sur la période initiale uniquement),
- un **nouveau modèle Mi** (entrainé sur la période jusqu'à la i -ème semaine).

3. Contrat de maintenance

Résultats



La période à partir de laquelle il faut ré-entrainer le modèle est de **11 semaines**
(soit **80 jours** pour un score **ARI de 80%)**

Segmentez des clients d'un site e-commerce

1 Préparation du jeu de données

Extraire les données caractérisant les clients à partir de la base de données Olist.

Nettoyage, sélection et création de variables, analyse exploratoire …

2 Modélisation et segmentation des clients

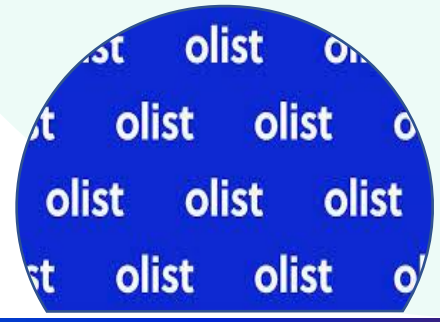
Segmenter les clients en fonction de leurs caractéristiques en utilisant les algorithmes de Machine Learning **non supervisés**.

Interpréter les segments obtenus d'un point de vue métier.

3 Simulation: contrat de maintenance

Analyser la stabilité temporelle de la segmentation pour évaluer une fréquence de maintenance

4 Conclusion



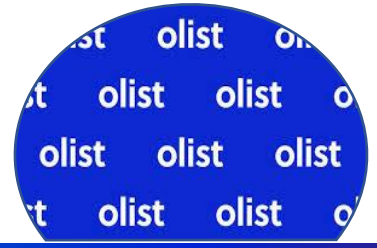
OPENCLASSROOMS

Segmentez des clients d'un site e-commerce

Conclusion

- Confirmer les compétences acquises en nettoyage et exploration des données
- Transformer les variables pertinentes d'un modèle d'apprentissage

- Mettre en place des **modèles d'apprentissage non supervisé** adapté au problème métier
- Evaluer les **performances** de ces modèles
- Adapter les **hyperparamètres** d'un algorithme d'apprentissage non supervisé afin de l'améliorer



Segmentez des clients d'un site e-commerce

Merci de votre attention